Responsible AI en security

Samaa Mohammad-Ulenberg l Platform voor informatiebeveiliging l 13.02.2024

Inhoudsopgave

Mijzelf en Alliander

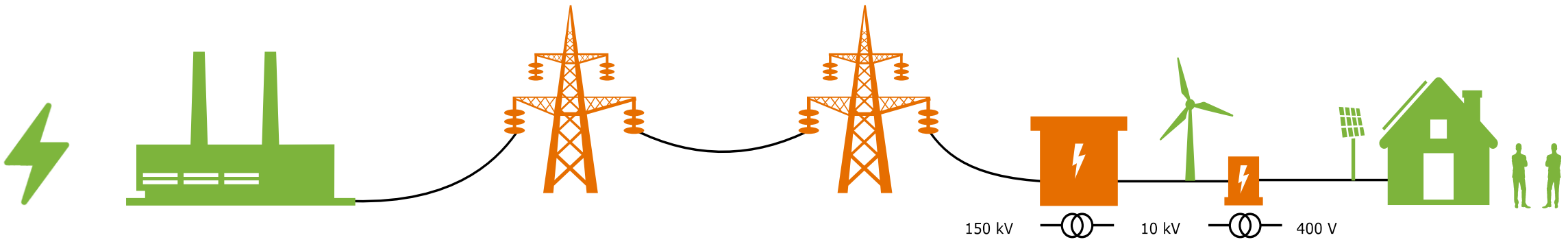
Introductie

Introductie

Samaa Mohammad-Ulenberg

Alliander	Digital ethics specialist
Algorithm Audit	Bestuurslid
Dutch AI Ethics Community	Oprichter & bestuurslid

Alliander develops and operates energy networks

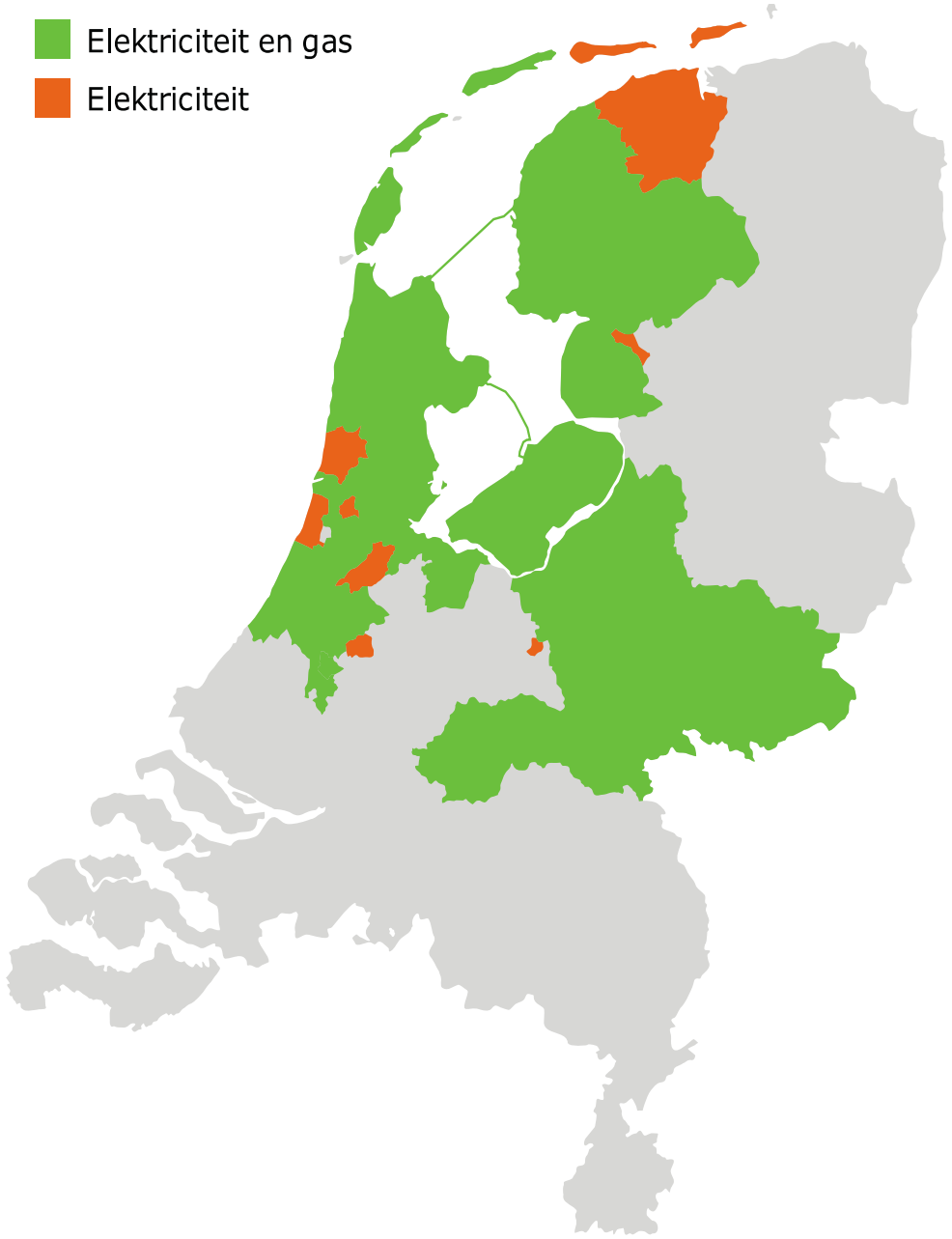


large-scale production

TSO
transmission system operator

DSO
distribution system operator

Illustrations



Electricity grid length
92,000 km

Gas grid length
42,000 km

Number of customer connections
5.8 million

Number of Employees
3700

Electricity outage duration
23.3 minutes

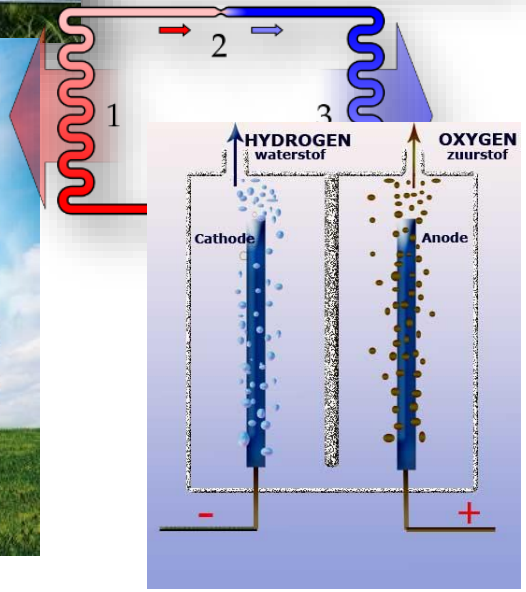
CO_2 emissions
205 kt

Net Revenue
€2.0 billion

Total assets
€9.4 billion

Grootste uitdagingen van Alliander op dit moment

Energietransitie



Maakbaarheid

Vaker vertraging bij aansluitingen Alliander

Door WOUTER VAN BERGEN
22 feb. 2018 in FINANCIEEL

AMSTERDAM - Klanten van netbeheerbedrijf Alliander lopen steeds vaker tegen vertraging op bij het verzorgen van nieuwe of betere aansluitingen. Dat vertelde Alliander-ceo Ingrid Thijssen bij de presentatie van zijn jaarresultaten donderdag.

Tekort aan technisch personeel blijft groeien

Gepubliceerd: 06 februari 2018 09:18
Laatste update: 06 februari 2018 10:21

Vacatures voor vakmensen in de techniek zijn steeds vaker moeilijk vervulbaar. Het aantal vacatures is in bijna twee jaar tijd met 16 procent gegroeid.

Door onze economieredactie

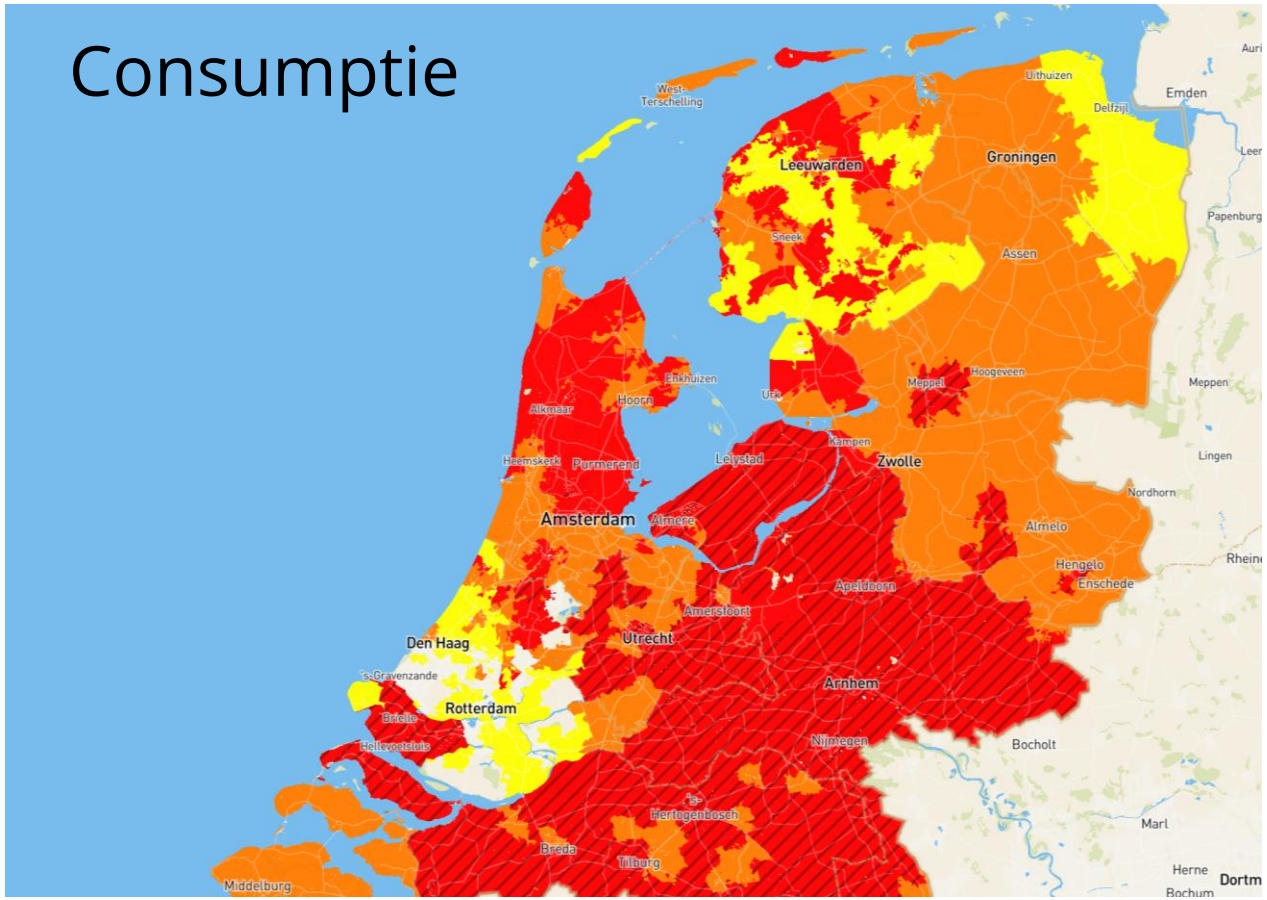
09 mrt 2023 om 08:48

680 reacties Delen

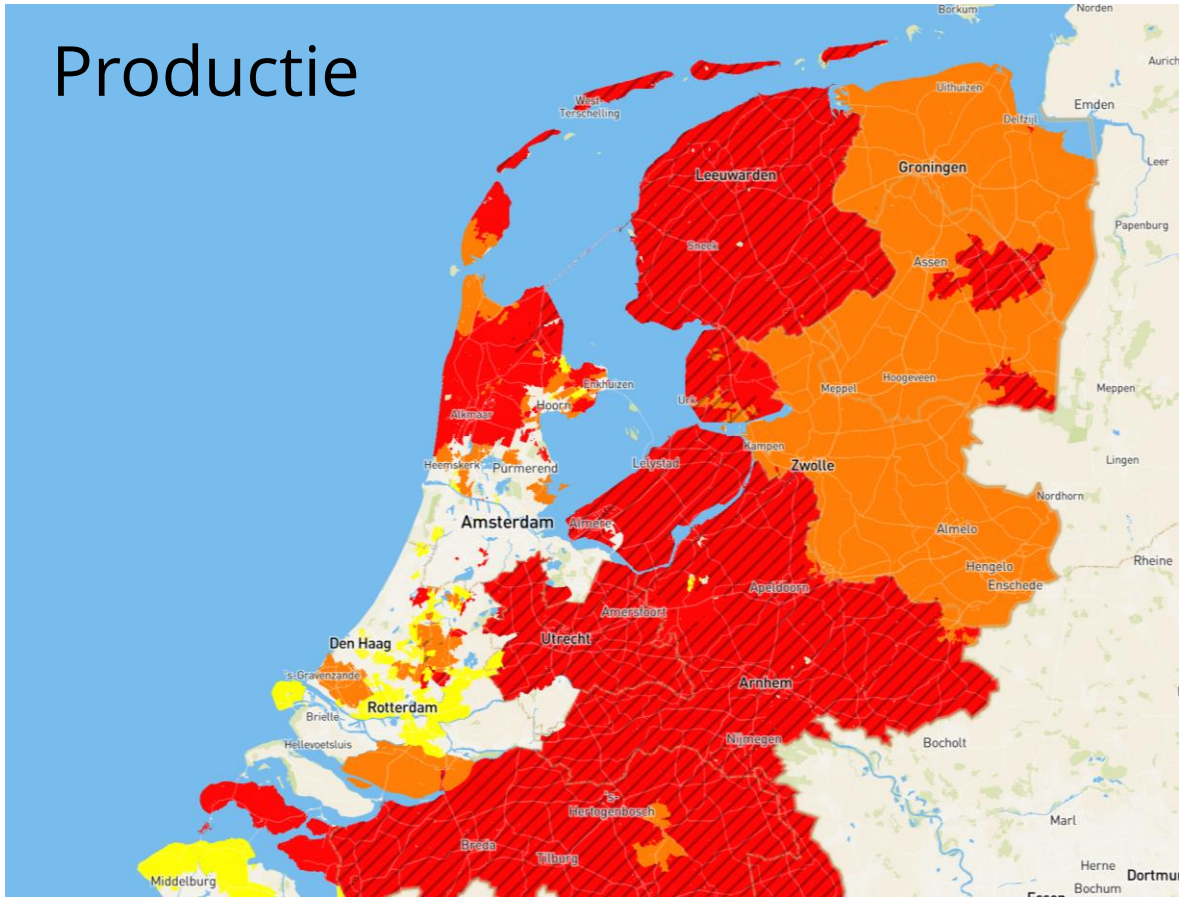
Netbeheerder Alliander kan de stijgende vraag naar elektriciteit nauwelijks aan. Het bedrijf heeft dan ook grote moeite om het netwerk in rap tempo uit te breiden nu consumenten en bedrijven meer elektrische auto's, zonnepanelen en warmtepompen hebben aangeschaft.

Effect: grootverbruik klanten kunnen niet aangesloten worden op het net

Consumptie



Productie



Betekenis van de kleurcodes

- ○ Transparant: Transportcapaciteit beschikbaar
- Geel: Beperkt transportcapaciteit beschikbaar
- Oranje: Voorlopig geen transportcapaciteit beschikbaar in afwachting van uitkomst van het congestiemanagement-onderzoek
- Rood: Geen transportcapaciteit beschikbaar: congestiemanagement kan niet worden toegepast

Met congestiemanagement:

- Transparant gearceerd: Transportcapaciteit beschikbaar o.b.v. toepassing congestiemanagement
- Geel gearceerd: Beperkt transportcapaciteit beschikbaar o.b.v. toepassing congestiemanagement
- Oranje gearceerd: Voorlopig geen transportcapaciteit beschikbaar in afwachting van het verdelen van het vrijgekomen vermogen over de wachtrij o.b.v. congestiemanagement. (het is nog onduidelijk of en hoeveel vermogen er beschikbaar komt voor nieuwe aanvragen die nog niet in de wachtrij staan)
- Rood gearceerd: Geen transportcapaciteit beschikbaar: de grenzen voor de toepassing van congestiemanagement zijn bereikt.

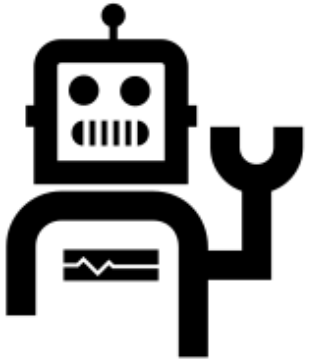
Waarom zetten we AI in bij Alliander?

Energietransitie & bereikbaarheid

- Efficiënter gebruik van het net, bv door meer inzicht, nieuwe contractvormen, slim laden, etc.
- Bepalen van meest efficiënte investeringen
- Proactief informeren van stakeholders over de status van het net

Maakbaarheid

- Automatiseren van repetitieve taken om de productiviteit van monteurs en engineers te verhogen.
- Betere kwaliteit van uitvoering en dataregistratie



Wat doet Alliander met AI?

AI

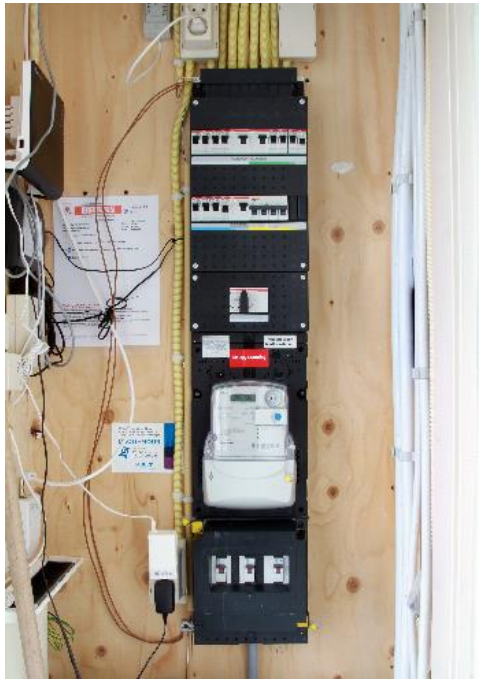
FAAM: Beeldherkenning in de meterkast

Foto-Automatisering Assets Meterkast

Klanten willen dat we hun stroomaansluiting versterken:

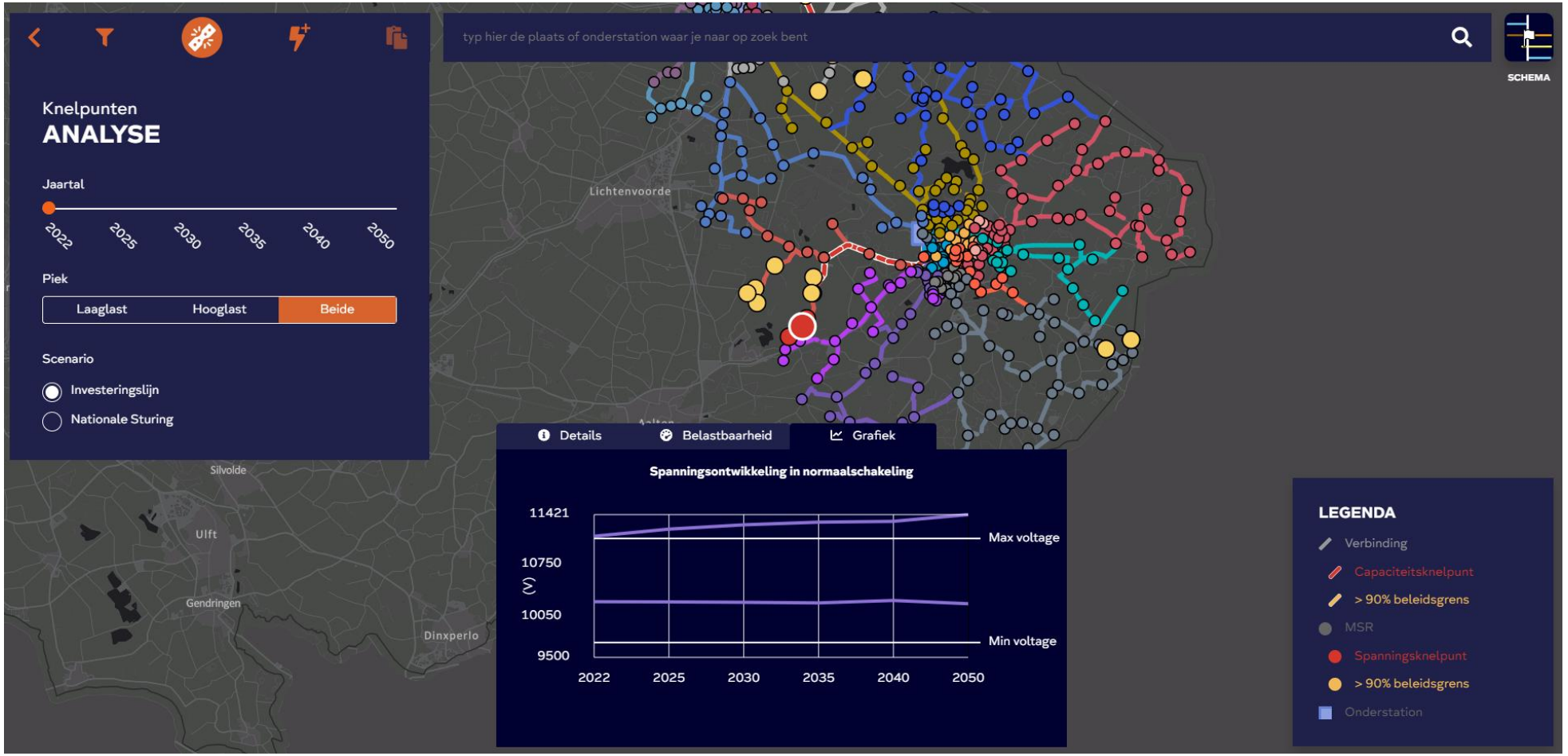
1. Klanten **uploaden foto** van hun meterkast in plaats van direct een afspraak te maken met monteur
2. **Beeldherkenning voert classificatie uit** in eenvoudige en moeilijke taken:

- **Makkelijk**: klant kan direct een afspraak inplannen met een monteur
- **Moeilijk**: de workflowmanager moet beslissen wat er moet gebeuren en wanneer met de klant overleggen.



Holonet en hololens

- Holonet en hololens



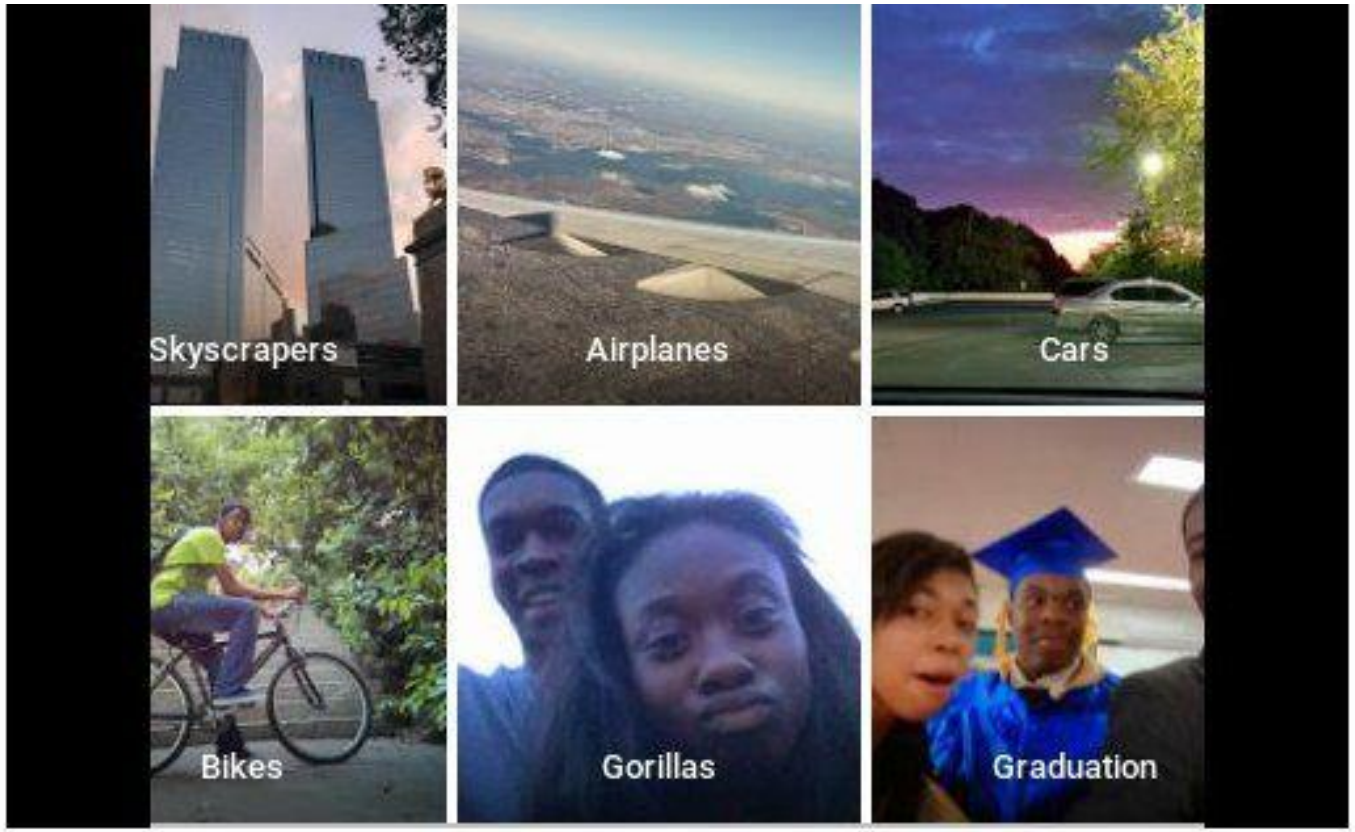
Verder:

- INPAI: Gevolgen van keuzes en gebeurtenissen te simuleren over 40 jaar
- AI for Energy Grids Lab
  - Onderzoeken van nieuwe AI oplossingen
- AI Acceleration team
  - Versnellen van AI ontwikkeling
- Digitaliseringsstrategie 2024-2030
- AI zou interessant kunnen zijn voor diverse bedrijfsprocessen, zoals:
  - Voorspellen van capaciteitsvraag, storingen en materiaalgebreken
  - Plannen van netuitbreidingen, onderhoud en supply chain
  - Ontwerpen van energienetten en software
  - Kennismanagement en digitale klantenservice

AI for Energy Grids Lab

Responsible AI

Responsible AI



alliander

Van dit...

...naar dit

COMPUTING

Racial Bias Found in a Major Health Care Risk Algorithm

Black patients lose out on critical care when systems equate health needs with costs

---

MIT apologizes, permanently pulls offline huge dataset that taught AI systems to use racist, misogynistic slurs

Machines Taught by Photos Learn a Sexist View of Women

Algorithms showed a tendency to associate women with shopping and men with shooting.

The algorithms that detect hate speech online are biased against black people

A new study shows that leading AI models are 1.5 times more likely to flag tweets written by African Americans as "offensive" compared to other tweets.

'The Computer Got It Wrong': How Facial Recognition Led To False Arrest Of Black Man

Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare

AI Bias Could Put Women's Lives At Risk – A Challenge For Regulators

Facebook's ad-serving algorithm discriminates by gender and race

Even if an advertiser is well-intentioned, the algorithm still prefers certain groups of people over others.

Apple's credit card is being investigated for discriminating against women

Customers say the card offers less credit to women than men

Google confirms it agreed to pay $135 million to two execs accused of sexual harassment

The $135 million was whittled down to $105 after one executive left to join Uber

YouTube Executives Ignored Warnings, Letting Toxic Videos Run Rampant

Proposals to change recommendations and curb conspiracies were sacrificed for engagement, staff say.

Inside Amazon: Wrestling Big Ideas in a Bruising Workplace

HOW FACEBOOK'S RISE FUELED CHAOS AND CONFUSION IN MYANMAR

The social network exploded in Myanmar, allowing fake news and violence to consume a country emerging from military rule.

Responsible AI

- Verantwoord ontwerpen, ontwikkelen en implementeren van AI
- AI is niet objectief, wij mensen hebben veel invloed op de resultaten van AI
- Dit kan onbedoeld leiden tot negatieve effecten
- Doel is om positieve impact te creëren en negatieve gevolgen te minimaliseren
- Grote topics binnen responsible AI o.a.: inclusiviteit, gelijke behandeling, privacy, uitlegbaarheid, veiligheid, duurzaamheid, transparantie
- Combinatie van compliance + ethiek -> wat is voorbij de wet, in deze context, het juiste om te doen?
- Het zetten van een eigen lat en kompas voor ethisch verantwoord AI
- Holistische en multidisciplinaire benadering technologie + cultuur + governance

Waarom responsible AI bij Alliander?

- Bijzondere maatschappelijke rol en verantwoordelijkheid
- Essentiele basis voor de digitale transformatie van Alliander
- Om technologie te kunnen bouwen waar wij en Alliander achter staan
- Om mensenrechten te kunnen beschermen
- Om risico’s en impact in kaart te brengen + te verantwoorden
- Om problematische uitkomsten te kunnen voorkomen
- Om een sterke basis te hebben om bij fouten wendbaarder te zijn
- De energietransitie zit al (zonder AI) boordevol ethische vraagstukken

- **Bovenal: AI Act en kritieke infra**

Onze benadering voor Responsible AI

- Oprichting **ethics team** medio 2023
- Gartner maturity assessment voor digital ethics, onze eerste stappen:
  - AI design principles
  - Ethisch risico assessment
  - AI governance framework
  - Ethische waarden
  - Algoritmeregister en quick scan voor risico classificatie algoritmes
  - Ethisch advies commissie
  - Data manifest
- **Iteratief proces om tot een volwassen systeem te komen**

AI Design Principles * Onder voorbehoud

Uitgangspunten bij de implementatie van AI: de waarden die wij belangrijk vinden bij de inzet van AI.

Fairness	Privacy & Security	Verklaarbaarheid en transparantie	Veiligheid en betrouwbaarheid
Inclusiviteit	Mensgerichtheid	Verantwoordelijkheid	Feedbackloop

1. Uitdaging: van principes naar concrete practises
2. ! Afwegingen zijn onvermijdelijk, toepassing is context-afhankelijk
3. Multidisciplinaire samenwerking tussen experts belangrijk om harmonie te creëren

Zoom in: security

Bescherming AI-specifieke dreigingen:

- Datavervuiling
- ‘Adversarial examples'
- Modeltekortkomingen zoals bias

AI systeem abuse VS onbedoelde negatieve gevolgen door mensen

Security 🤝 Responsible AI maakt een sterk AI Red team approach!

Security feedback loops essentieel

Onze benadering is holistisch, denk aan:

Governance	Technologie	Cultuur
• Impact assessments voor mensenrechten en algoritmes (IAMA) • Algoritmeregister • Stakeholder participatie • Accountability • Risico documentatie en management • Feedbackloops • Audits • etc	• AI Design Principles • Model, variabele en feature selectie • Responsible AI tooling • Data kwaliteit en representatie • Data beheer • Documentatie • Security maatregelen (data vervuiling, model manipulatie, etc) • etc	• Cultuurverandering • Ethische waarden en standaarden en institutionalisering in besluitvorming • Diversiteit teams • Bewustzijn • Kennisopbouw • Pro-activiteit • Mindset context-sensitief ethisch reflecteren • etc

Casus

Belang deliberatie en multidisciplinaire kijk op AI casuïstiek

Casus Algorithm Audit

- Algorithm Audit is een europees kennisplatform voor AI bias testing en normatieve ethische standaarden.
- Ethisch advies commissies met diverse mensen en disciplines om knopen door te hakken
- Belang multidisciplinaire deliberatie

Case study: Algoritmische risicoprofilering om fraude te voorkomen bij achteraf betalen

Een e-commerce platform biedt een service aan dat het voor klanten mogelijk maakt om achteraf te kunnen betalen

>

Sommige klanten kopen wel, maar betalen nooit

- Jaarlijks gaan ongeveer 1.550 van de 25.000 betalingen in gebreke (6,2%)
- Dit resulteert in een verlies van €110k, wat 4,4% van de totale omzet (€25M) bedraagt

>

Historische data kan gebruikt worden om achter de eigenschappen van wanbetalers te komen

Beschikbare databronnen:

- Gebruikerseigenschappen: bijv. adres, contactgegevens
- Data uit gedrag: bijv. betaalgeschiedenis, account leeftijd, platform interactie
- Metadata van mobiele app, bijv. type telefoon, type simkaart

Case study: Algoritmische risicoprofilering om fraude te voorkomen bij achteraf betalen (vervolg)

Data analyseren om inzicht te krijgen op wat voor type klanten een grote kans hebben om niet te betalen

Resultaat: klanten die gemiddeld:
- voor meer dan €100 bestellen; of
- een bepaalde type simkaart hebben;

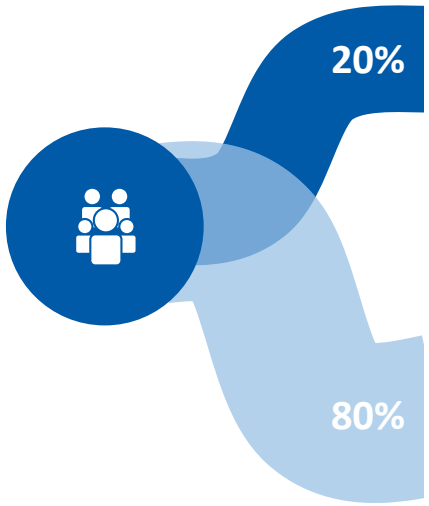
hebben een grotere kans om niet te betalen

>

Een filter toepassen om te bepalen welke klanten in aanmerking komen om achteraf te kunnen betalen

Filter:
- bestel voor meer dan €100; of
- gebruik van een bepaald type simkaart;

achteraf betalen niet mogelijk: betaal direct



Risico → Betaal direct

Geen risico → Komt in aanmerking voor achteraf betalen

A/B-testen tonen aan dat filteren het aantal wanbetalingen voor achteraf betalen met 50% vermindert, d.w.z. een afname van 6,2% naar 3,1%, wat jaarlijks ~€59k bespaart

Ethisch dilemma: Proxy discriminatie (gelijke behandeling) vs voorspellende kracht (nauwkeurigheid)

Proxy discriminatie

Sommige soorten simkaarten zijn sterk gecorreleerd met bepaalde demografische groepen, wat het risico op proxy (indirecte) discriminatie creëert

Voorspellend vermogen

Statistische analyse geeft aan dat 'type simkaart' voorspellende kracht heeft om potentiële fraudulente klanten te identificeren

Het weglaten van 'type simkaart' uit het filter resulteert in minder nauwkeurige resultaten:

Zonder filter	Met filter	
6.2% wanbetaling	3.1% wanbetaling	Met sim
	4.9% wanbetaling	Zonder sim

~€40k verschil

Hoofdvraag

In welke mate en onder welke omstandigheden is het ethisch verantwoord om de variabele 'type simkaart' te gebruiken in voorspellingsmodellen om achteraf-betaaldiensten te blokkeren voor specifieke klanten die als risicovol worden geclassificeerd?

Vraag voor het publiek

1. Welke security risico's zijn er als variabele 'type simkaart' wordt gebruikt in het voorspellingsmodel om after-pay diensten te blokkeren? Welke security risico's zijn er als deze variabele wordt weggelaten?
2. Stel we gaan uit van AI design principle 'transparantie' en het algoritme moet duidelijkheid bieden welke variabelen zijn gebruikt in de vorming van de beslissing: welke risico's voor abuse zien jullie? Hoe staan jullie tegenover security vs transparantie?
3. Andere gedachten/perspectieven op deze case?

Advies

De auditcommissie raadt het gebruik van de variabele 'type simkaart' af

Type simkaart waarschijnlijk proxyvariabele

Overweeg alternatieven, zoals hoofdoorzaak wanbetaling onderzoeken of publiek voor achteraf betalen wijzigen.

Bedankt voor jullie aandacht!

Samaa Mohammad-Ulenberg

Samaa.mohammad-ulenberg@alliander.com

LinkedIN: @Samaam